# Studi Perbandingan Metode *Naive Bayes* dan *Linear Discriminant Analysis* Untuk Permasalahan Klasifikasi

Muhammad Irsyad Razan[1], Muhammad Fatchan[1], Regan Agam[1], Gita Suryani[1], Reza Hadi Asykari[1], Aldi Darma[1]
[1]Teknik Elektro, Fakultas Teknik
Universitas Sriwijaya
Palembang, Indonesia
Penulis korespodensi: irsyad301200.ir@gmail.com

***Abstrak*— Pemanfaatan teknologi tidak bisa lepas dari kehidupan manusia saat ini. Berbagai macam teknologi tersebut salah satunya seperti komputer atau laptop sangat penting dalam membantu urusan pengolahan data. Ada banyak metode yang dapat digunakan dalam melakukan pengolahan data. Salah satu metodenya disebut dengan klasifikasi. Permasalahan yang dibahas pada kasus kali ini yaitu membandingkan klasifikasi LDA dan *Naive Bayes* berdasarkan jumlah data, fitur dan kelas pada dataset yang berbeda dengan tujuan untuk menganalisis tingkat keakurasian pada setiap metode klasifikasi, kemudian akan dilihat metode klasifikasi yang lebih akurat berdasarkan bentuk dan jenis data yang telah diklasifikasi sebelumnya. Pengklasifikasian data menggunakan algoritma metode LDA dan *Naive Bayes* dengan keluarannya berbentuk matriks kekeliruan (*Confusion Matrix*) dan tingkat keakuratan agar dapat dibandingkan diantara klasifikasi tersebut. Pada *dataset obesity*, metode klasifikasi LDA mampu memprediksi 391 dari 423 data dengan tingkat akurasi 92,2%, sedangkan metode klasifikasi *Naive Bayes* hanya dapat memprediksi 307 dari 528 data dengan tingkat akurasi 58,1%. Pada *dataset wine*, terdapat 35 data prediksi pada klasifikasi LDA dengan tingkat akurasi 97,2% dan 44 data prediksi pada klasifikasi *Naïve Bayes* dengan tingkat akurasi 97,7% serta hanya satu data yang tidak dapat diprediksi dari setiap klasifikasi. Dengan demikian, dapat disimpulkan bahwa klasifikasi LDA memiliki tingkat keakuratan yang lebih baik dibandingkan klasifikasi *Naive Bayes*.**

***Kata kunci*—klasifikasi, LDA, naive bayes, tingkat akurasi**

***Abstract*— The use of technology cannot be separated from human life today. Various kinds of technology, one of which is a computer or laptop, which is very important in helping with data processing. There are many methods that can be used in data processing. One method is called classification. The problem discussed in this case is to compare the LDA and Naive Bayes classifications based on the amount of data, features and classes in the dataset with the aim of analyzing the level of accuracy in each classification method, then a more accurate classification method will be seen based on the form and type of data that has been collected. previously classified. Classification of data using LDA and Naive Bayes methods with the output in the form of a matrix (Confusion Matrix) and the level of accuracy so that it can be compared between these classifications. In the obesity dataset, the LDA classification method was able to predict 391 of 423 data with an accuracy rate of 92.2%, while the Naive Bayes classification method could only predict 307 of 528 data with an accuracy rate of 58.1%. In the wine dataset, there are 35 predictive data on the LDA classification with an accuracy of 97.2% and 44 predictive data on the Naïve Bayes classification with an accuracy of 97.7% and only one data that cannot be predicted from each classification. Thus, it can be said that the LDA classification has better accuracy than the Naive Bayes classification.**

***Keywords—classification, LDA, naive bayes, level of accuracy***

## I. PENDAHULUAN

Zaman sekarang, pemanfaatan teknologi tidak bisa lepas dari kehidupan manusia saat ini. Bisa dilihat dari perusahaan besar maupun kecil yang terus bersaing demi mendapatkan terobosan terbaru untuk mewujudkan teknologi yang lebih canggih dari sebelumnya. Perangkat seperti komputer ataupun laptop merupakan salah satu teknologi yang sangat penting saat ini dalam mengolah suatu data. Dalam pengolahan data ini tentu saja ada banyak metode yang dapat digunakan secara umum oleh banyak orang. Seiring dengan perkembangan teknologi, metode tersebut terus-menerus berkembang hingga saat ini. Salah satu metodenya itu disebut dengan klasifikasi.

Klasifikasi merupakan suatu metode pengelompokan yang sistematis terhadap sejumlah objek, gagasan, buku atau benda-benda lain ke dalam suatu kelas atau golongan berdasarkan ciri-ciri yang sama. Dalam klasifikasi ini dapat dilakukan dengan cara manual atau bisa juga dengan bantuan teknologi. Akan tetapi, pada saat ini banyak orang menggunakan bantuan teknologi dalam mengklasifikasikan suatu data dengan tujuan untuk memperoleh kemudahan serta meningkatkan keefisienan dalam bekerja. Adapun berbagai metode klasifikasi dengan bantuan teknologi seperti *Naive Bayes*, *support Vector Machine*, LDA, *fuzzy*, dan lain-lain.[1]

Dari banyaknya metode klasifikasi tersebut, percobaan pada kedua data (*wine* dan *obesity*) ini menggunakan dua metode klasifikasi yaitu *Naive Bayes Classifier* dan *Linear Discriminant Analysis* (LDA) yang berfungsi untuk mengklasifikasi atau menghimpun data dalam bentuk suatu kelompok. Permasalahan yang dibahas pada kasus kali ini ialah membandingkan klasifikasi LDA dan *Naive Bayes* berdasarkan jumlah data, fitur dan kelas pada dataset yang berbeda dengan bertujuan untuk menganalisis tingkat keakurasian pada setiap metode klasifikasi

yang berdasarkan pada jenis data, jumlah banyaknya data maupun banyaknya variabel data tersebut. Dan dari hasil percobaan ini, nantinya akan dapat dilihat metode klasifikasi mana yang lebih akurat berdasarkan bentuk dan jenis data yang telah diklasifikasi sebelumnya.

## II. METODOLOGI PENELITIAN

### *A. Pengumpulan Data*

Tahap awal dari metode penelitian ini adalah pencarian *dataset* yang kami dapatkan dari *website* UCI *Machine Learning* (https://archive.ics.uci.edu/ml/index.php). Kemudian selanjutnya kami mengambil data set *Wine* dan *Obesity*. Pada dataset *obesity*, terdapat 2111 data, 16 fitur, dan 7 klasifikasi. Ke 16 fitur tersebut merupakan variabel bebas yang dapat mempengaruhi klasifikasi terhadap *dataset obesity*. Adapun klasifikasi yang akan digunakan pada *dataset* ini berupa *Normal Weight*, *Insufficient Weight*, *Overweight* Level 1, *Overweight* Level 2, *Obesity* Type 1, *Obesity* Type 2, dan *Obesity* Type 3. Pada *dataset Wine,* terdapat 176 data, 13 fitur dan 3 klasifikasi. Adapun klasifikasi yang akan digunakan pada *dataset* ini yaitu *wine* 1, *wine* 2 dan *wine* 3.

### *B. Perumusan Data Klasifikasi*

#### *1) LDA*

Fungsi diskriminan dibentuk untuk pemisahan kelompok. Hasil dari analisis diskriminan ini diperoleh suatu fungsi yang digunakan untuk mengelompokkan pengamatan ke dalam suatu kelas, yang kemudian disebut fungsi diskriminan. Matriks kovarians intrakelas ($S_w$) pada persamaan (1) dan matriks kovarians kelas ($S_B$) melalui persamaan (2) didefinisikan sebagai berikut [2]:

$$S_w = \sum_{i=1}^{c} \sum X_k \in X_i \, (X_k - \mu_i)(X_k - \mu_i)^T \quad (1)$$

$$S_B = \sum_{i=1}^{c} N_i \, (\mu_i - \mu)(\mu_i - \mu)^T \quad (2)$$

Agar matriks kovarians dalam kelas ($S_w$) dapat diminimalkan sedangkan matriks kovarians antar kelas ($S_B$) maksimal, akan dicari vektor eigen (V) pada persamaan (3) untuk menskalakan persamaan hingga maksimum:

$$V = \frac{\det (V S_B V^T)}{\det (V S_w V^T)} \quad (3)$$

Oleh karena itu menghasilkan solusi dengan persamaan (4):

$$S_B V = \lambda S_w V \quad (4)$$

Selanjutnya, nilai eigen (λ) dan nilai eigen (V) diperoleh dari persamaan matriks kovarians menghasilkan persamaan (5), yaitu:

$$Cov = S_B {S_w}^{-1} \quad (5)$$

Setelah diketahui eigen vector, maka nilai karakteristik LDA dapat dihitung menggunakan persamaan (6):

$$F_x = \sum_{i=1}^{k} (X_i - \mu)^T \times V \quad (6)$$

Pencarian jarak menggunakan jarak Euclidean dengan persamaan (7):

$$E(A, B) = \sum_{i=1}^{N} \sqrt{(A_i - B_i)^2} \quad (7)$$

#### *2) Naive Bayes*

Proses klasifikasi memerlukan beberapa petunjuk untuk menentukan jenis yang sesuai untuk sampel yang dianalisis. Oleh karena itu, metode *Naive Bayes* dapat ditulis melalui persamaan (8):

$$Posterior\ Probability = Likelihood\ x\ \frac{Class\ Prior\ Probability}{Predictor\ Prior\ PRobabilty} \quad (8)$$

Langkah-langkah algoritma *Naive Bayes* [3]:

1. Pembacaan data pelatihan
2. Perhitungan jumlah kelas.
3. Perhitungan jumlah kasus yang sesuai dengan dengan (probabilitas).
4. Kalikan semua nilai yang dihasilkan dengan data baru yang dicari lapisan. Hasil pemilihan kelas dibandingkan dengan nilai tertinggi.

Prediksi Bayes didasarkan pada Teorema Bayes dengan persamaan umum (9) dan persamaan (10) berikut:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (9)$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n|c) \times P(c) \quad (10)$$

dimana

- P(c|x) adalah probabilitas posterior kelas (target) yang diberikan prediktor (atribut).
- P(c) adalah peluang kelas sebelumnya.
- P(x|c) adalah peluang yang merupakan peluang dari kelas yang diberikan oleh prediktor.
- P(x) adalah probabilitas sebelumnya dari prediktor.

### *C. Algoritma*

Adapun untuk pengklasifikasian data kami menggunakan algoritma metode LDA dan *Naive Bayes* yang keluarannya berbentuk *confusion matrix* dan tingkat keakuratan agar dapat dibandingkan diantara klasifikasi tersebut. Pada penelitian ini, kami menggunakan bahasa pemrograman *Python* dengan *software Visual Studio Code*. Berikut adalah alur pemrograman klasifikasi:

#### *1) Linear Discriminant Analysis (LDA)*

```
1  #Library dan Prosedur LDA
2  import numpy as np
3  import matplotlib.pyplot as plt
4  import pandas as pd
5  from sklearn.preprocessing import LabelEncoder as LE
6  from sklearn.model_selection import train_test_split
7  from sklearn.preprocessing import StandardScaler
8  from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
9  from sklearn.linear_model import LogisticRegression
```

Gambar 1. Alur algoritma LDA

Keterangan pada Gambar 1 yaitu:
Line 2 : *library numpy* berfungsi untuk melakukan komputasi matriks
Line 3 : pada *library matplotlib* dengan modul *pyplot* digunakan untuk *plotting* atau membuat plot
Line 4 : *library pandas* berfungsi untuk mengambil dataset yang digunakan
Line 5 : pada *library sklearn* dengan modul *preprocessing* diimport *Label Encoder* untuk melakukan proses pemisalan dari bentuk data *string* menjadi integer/angka untuk beberapa data dibeberapa fitur
Line 6 : pada *library sklearn* dengan modul *model_selection* diimport *train_test_split* untuk melakukanproses pembagian data kedalam bentuk data *training* dan data *testing* dengan komposisi 80:20
Line 7 : pada *library sklearn* dengan modul *preprocessing* diimport *Standar Scaler* untuk melakukan fitur *scalling* pada *x_train dan x_test*
Line 8 : pada *library sklearn* dengan modul *discriminant_analysis* diimport *Linear Discriminant Analysis* untuk menjalankan algoritma LDA
Line 9 : pada *library sklearn* dengan modul *linear_model* diimport *Logistic Regression* sebagai klasifikasi tambahan untuk membantu algoritma LDA

#### *2) Naive Bayes*

```
1  # Library dan Prosedur LDA
2  import numpy as np
3  import matplotlib.pyplot as plt
4  import pandas as pd
5  from sklearn.preprocessing import LabelEncoder as LE
6  from sklearn.model_selection import train_test_split
7  from sklearn.preprocessing import StandardScaler
8  from sklearn.naive_bayes import GaussianNB
```

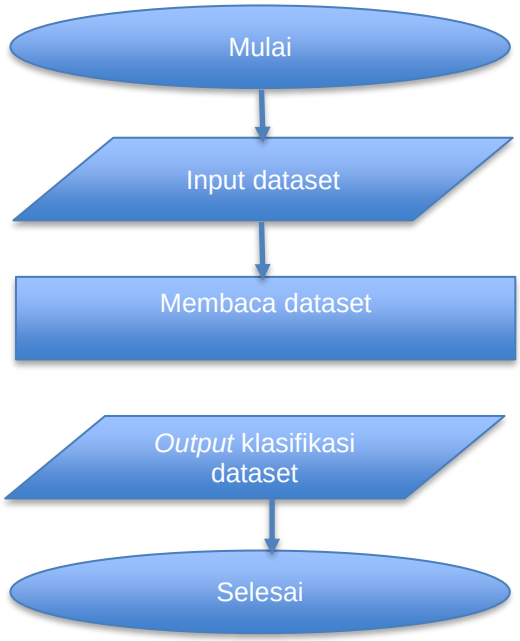
Gambar 2. Alur algoritma *naive bayes*

Keterangan pada Gambar 2 ialah:
Line 2 : *library numpy* berfungsi untuk melakukan komputasi matriks
Line 3 : pada *library matplotlib* dengan modul *pyplot* digunakan untuk *plotting* atau membuat plot
Line 4 : *library pandas* berfungsi untuk mengambil dataset yang digunakan

Line 5 : pada *library sklearn* dengan modul *preprocessing* diimport *Label Encoder* untuk melakukan proses pemisalan dari bentuk data *string* menjadi integer/angka untuk beberapa data dibeberapa fitur
Line 6 : pada *library sklearn* dengan modul *model_selection* diimport *train_test_split* untuk melakukan proses pembagian data kedalam bentuk data *training* dan data *testing* dengan komposisi 80:20
Line 7 : pada *library sklearn* dengan modul *preprocessing* diimport *Standar Scaler* untuk melakukan fitur *scalling* pada *x_train dan x_test*
Line 8 : pada *library sklearn* dengan modul *naive_bayes* diimport *GaussianNB* untuk menjalankan algoritma *Naive Bayes* dengan metode *GaussianNB*

### *D. Flowchart Penelitian*

Penelitian ini menggunakan *flowchart* penelitian seperti pada Gambar 3:




Gambar 3. *Flowchart* penelitian

## III. HASIL DAN ANALISIS

### *A. Hasil*

#### *1) Confusion Matrix*

TABEL I. CONFUSION MATRIX

| | **LDA** | ***Nalve Bayes*** |
|---|---|---|
| Wine | $\begin{bmatrix}14 & 0 & 0\\ 0 & 15 & 0\\ 0 & 1 & 6\end{bmatrix}$ | $\begin{bmatrix}15 & 1 & 0\\ 0 & 20 & 0\\ 0 & 0 & 9\end{bmatrix}$ |
| Obesity | $\begin{bmatrix}63 & 2 & 0 & 0 & 0 & 0 & 0\\ 6 & 41 & 0 & 0 & 0 & 10 & 0\\ 0 & 0 & 66 & 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 53 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 69 & 0 & 0\\ 0 & 3 & 0 & 0 & 0 & 53 & 1\\ 0 & 0 & 2 & 0 & 0 & 7 & 46\end{bmatrix}$ | $\begin{bmatrix}77 & 1 & 0 & 0 & 0 & 0 & 0\\ 59 & 4 & 1 & 0 & 1 & 3 & 1\\ 0 & 1 & 67 & 14 & 1 & 1 & 2\\ 0 & 0 & 12 & 59 & 0 & 0 & 2\\ 0 & 1 & 0 & 0 & 87 & 0 & 0\\ 26 & 4 & 31 & 0 & 0 & 6 & 1\\ 8 & 5 & 34 & 9 & 0 & 3 & 7\end{bmatrix}$ |

Tabel 1 adalah kumpulan *confusion matrix* dari hasil klasifikasi dataset *Wine* dan *Obesity*. Pada dataset *wine* hanya terdapat satu data dari setiap klasifikasi yang tidak dapat diprediksi, namun data prediksi untuk setiap kategori berbeda jumlahnya, pada klasifikasi LDA terdapat 35 data prediksi. pada klasifikasi *Naive Bayes* memiliki 44 data prediksi. Perbedaan hasil dari matriks konfusi ini memungkinkan untuk membedakan keakuratan suatu metode klasifikasi.

Pada dataset *obesity,* metode klasifikasi LDA mampu memprediksi 391 dari 423 data, sedangkan metode klasifikasi *Naive Bayes* hanya dapat memprediksi 307 dari 528 data. Dapat kita lihat disini bahwa metode klasifikasi *Naive Bayes* memiliki banyak data yang tidak dapat diprediksi. Hal ini dimungkinkan karena dipengaruhi oleh jumlah data, fitur dan klasifikasi pada dataset *obesity*. Dari keseluruhan *confusion matrix* dapat kita analisis bahwa Metode *Naive Bayes* akan menghasilkan data yang lebih banyak secara keseluruhan dibandingkan dengan metode LDA. Akan tetapi, Metode LDA memiliki jumlah data tidak terprediksi lebih sedikit dibandingkan Metode *Naive Bayes*.

*2) Tingkat Keakurasian*


Gambar 4. Tingkat keakurasian

Gambar 4 merupakan perbandingan tingkat keakuratan dari metode klasifikasi LDA dan *Naive Bayes* menggunakan dua dataset yang berbeda. Tingkat keakuratan ini adalah hasil dari perhitungan data yang diprediksi dan tidak terprediksi pada *confusion matrix*. Dapat kita lihat bahwa klasifikasi *Naive Bayes* memiliki tingkat keakuratan yang rendah untuk dataset dengan  data, fitur dan klasifikasi yang banyak dikarenakan banyak data yang tidak dapat diprediksi pada proses *confusion matrix*. Sedangkan untuk klasifikasi LDA memiliki tingkat keakuratan yang lebih tinggi karena dapat mengecilkan kemungkinan data yang tidak terprediksi.

## IV. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, banyaknya jumlah data, fitur dan klasifikasi dapat mengetahui keakuratan pada saat proses klasifikasi data. Baik klasifikasi LDA maupun klasifikasi *Naive Bayes* apabila dataset yang dipakai semakin sedikit, maka tingkat keakurasian pada proses klasifikasi data akan semakin tinggi. Hasil yang telah kami dapat dari klasifikasi LDA dan *Naive Bayes* diatas, Klasifikasi LDA mampu mempertahankan keakuratannya pada dataset kecil maupun besar sedangkan Klasifikasi *Naive Bayes* tidak mampu mempertahankan keakuratannya dalam mengolah dataset besar. Sehingga dapat diambil kesimpulan bahwa klasifikasi LDA tingkat keakuratannya lebih baik dibandingkan klasifikasi *Naive Bayes*.

## DAFTAR PUSTAKA

[1] R., Salsabila Miftah, “Macam-Macam Algoritma Klasifikasi Machine Learning Yang Penting Untuk Diketahui,” https://www.dqlab.id/macam-algoritma-klasifikasi-machine-learning-yang-penting-untuk-diketahui, diakses pada tanggal 26 Oktober 2021. 2021.

[2] Sari, R. P., Rosiani, U. D., dan Syulistyo, A. R., “Implementasi Metode Linear Discriminant Analysis Untuk Deteksi Kematangan Pada Buah Stroberi,” (Doctoral Dissertation, Teknologi Informasi). 2020.

[3] Fanani, M. R. “Algoritma Naïve Bayes Berbasis Forward Selection Untuk Prediksi Bimbingan Konseling Siswa,” Jurnal Disprotek, 11. 2020.